

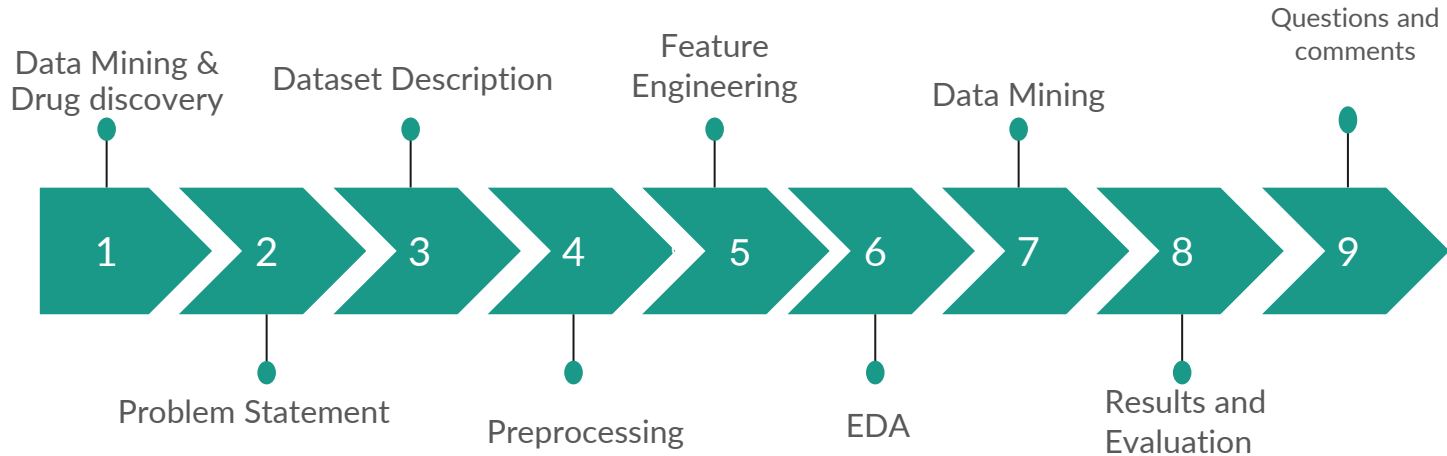


# Data Mining Applications in The Healthcare field

AI-based Quantitative structure Activity relationship  
study (QSAR) for Alzheimer's disease

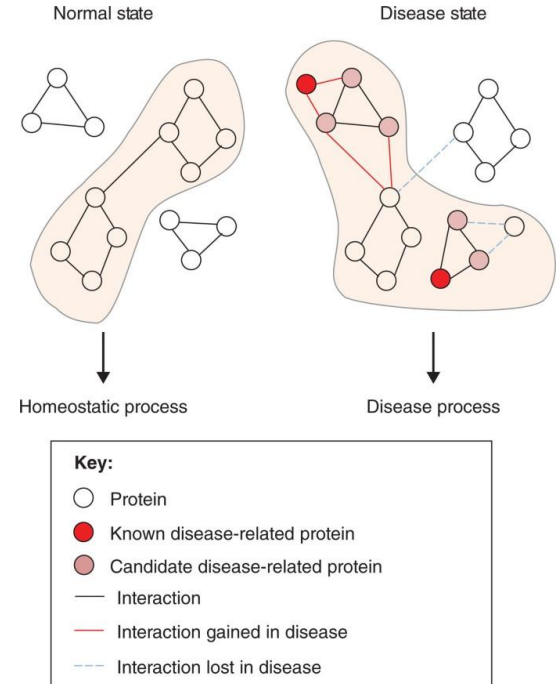
Data Mining Course –Ajman University

# List of Content

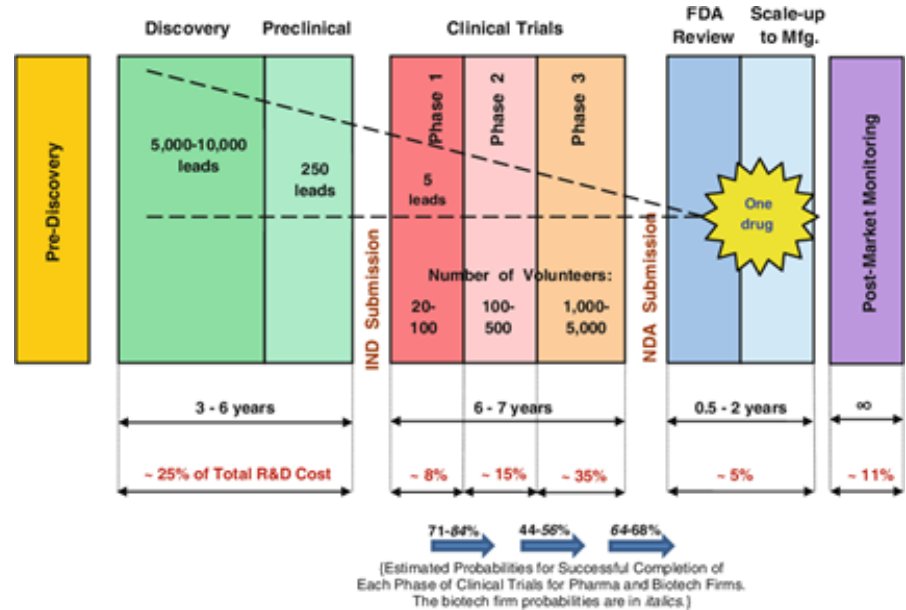


# Drug Discovery & Data Mining :Story behind it

- Also called **Quantitative Structure Activity Relationship (QSAR)**
- Based on hypotheses at what is the targeted protein
- search for the molecule that changes the functionality of target protein



# Drug Discovery Development Process



---

# Problem Statement

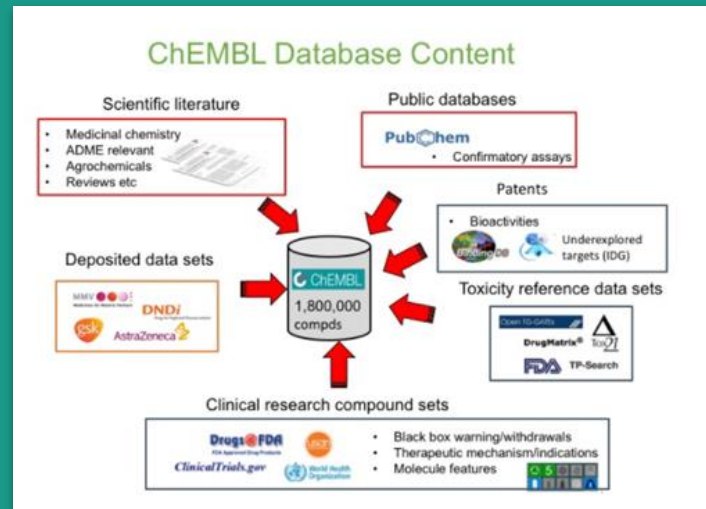
*The process of drug discovery is a very long process that can take years of research, testing phases and getting the approval from Federal to be available to public for use. Utilizing Machine learning Algorithm, we aim to automate the routine work in the drug discovery laboratories of manually observing the activities of the target protein over the span of years and a load of biological calculations and analytical statistics.*

# Database Description

Database : ChEMBL

Disease: Alzheimer

Target protein : Amyloid Beta A4 protein



# Dataset Description

```
#print(df.shape)
print(df.info())

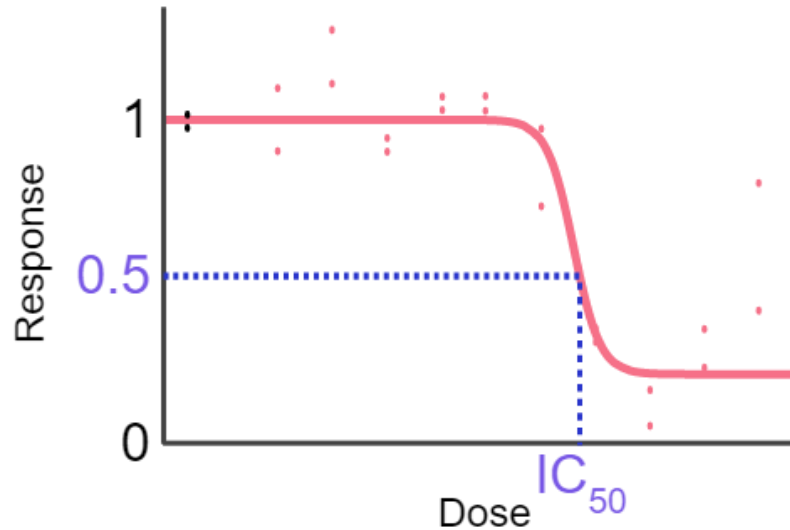
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1245 entries, 0 to 1244
Data columns (total 45 columns):
#   Column                Non-Null Count  Dtype
---  -
0   activity_comment       284 non-null   object
1   activity_id            1245 non-null  int64
2   activity_properties     1245 non-null  object
3   assay_chembl_id        1245 non-null  object
4   assay_description      1245 non-null  object
5   assay_type             1245 non-null  object
6   assay_variant_accession 0 non-null     object
7   assay_variant_mutation 0 non-null     object
8   bao_endpoint           1245 non-null  object
9   bao_format             1245 non-null  object
10  bao_label              1245 non-null  object
11  canonical_smiles       1245 non-null  object
12  data_validity_comment   37 non-null    object
13  data_validity_description 37 non-null    object
14  document_chembl_id      1245 non-null  object
15  document_journal        1088 non-null  object
16  document_year           1245 non-null  int64
17  ligand_efficiency       899 non-null   object
18  molecule_chembl_id      1245 non-null  object
19  molecule_pref_name      128 non-null   object
20  parent_molecule_chembl_id 1245 non-null  object
```

```
21  pchembl_value           936 non-null   object
22  potential_duplicate     1245 non-null  bool
23  qudt_units              1115 non-null  object
24  record_id               1245 non-null  int64
25  relation                 1113 non-null  object
26  src_id                  1245 non-null  int64
27  standard_flag           1245 non-null  bool
28  standard_relation       1113 non-null  object
29  standard_text_value     0 non-null     object
30  standard_type           1245 non-null  object
31  standard_units          1117 non-null  object
32  standard_upper_value    0 non-null     object
33  standard_value          1117 non-null  object
34  target_chembl_id        1245 non-null  object
35  target_organism         1245 non-null  object
36  target_pref_name        1245 non-null  object
37  target_tax_id           1245 non-null  object
38  text_value              0 non-null     object
39  toid                    0 non-null     object
40  type                    1245 non-null  object
41  units                   1147 non-null  object
42  uo_units                1115 non-null  object
43  upper_value             4 non-null     object
44  value                   1117 non-null  object
dtypes: bool(2), int64(4), object(39)
memory usage: 420.8+ KB
None
```

Shape = (1245, 45)

2 important attributes: canonical\_smiles (obj), standard\_value(obj)

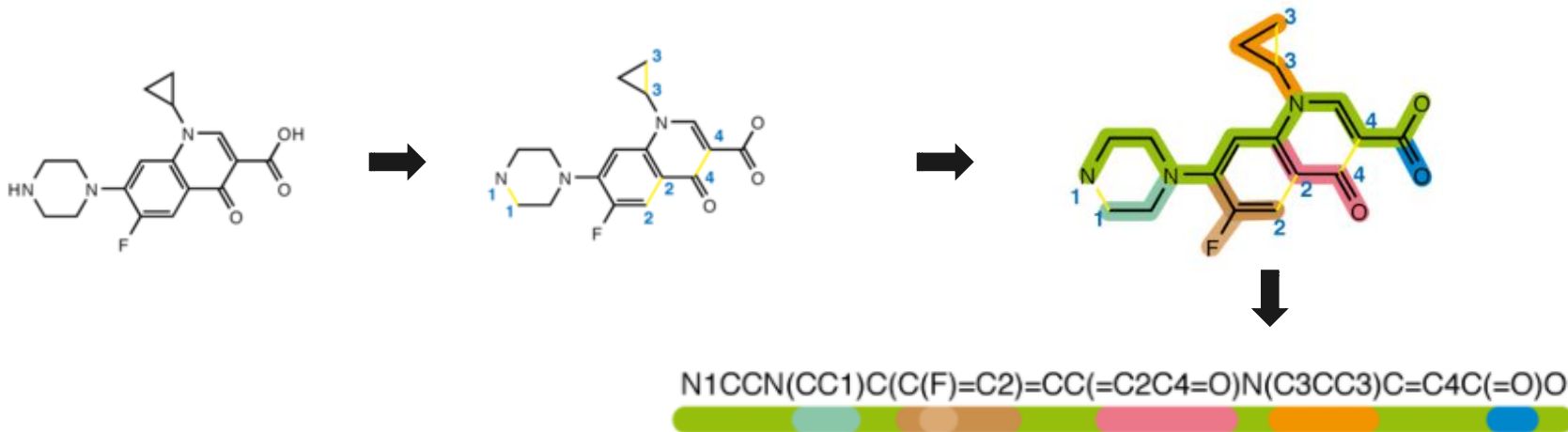
# IC<sub>50</sub> and pIC<sub>50</sub>



- Half maximum inhibition concentration : a measure of the potency of a substance in inhibiting a specific biological or biochemical function.
- pIC<sub>50</sub> is negative log of IC<sub>50</sub>

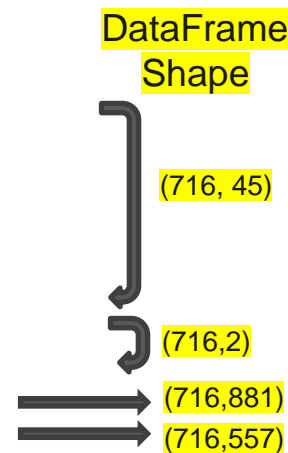


# Smiles (Simplified Molecular input Line-Entry System)



# Data Preprocessing

1. Data Preprocessing
  - a. Converted standard\_value to float
  - b. Drop rows that have *canonical\_smiles* or \and *standard\_value* as na (missing data)
  - c. Drop rows that have duplicate smiles notation (duplicated data)
  - d. Discretization: Create bioactivity **class** attribute and remove intermediate bioactive molecules (discretization)
  - e. Simplify the smiles notation and remove non-bonded elements (remove noise)
  - f. Normalize the standard value by taking the negative logarithmic value of IC50 == pIC50 (normalization)
  - g. Select The most meaningful features for our experiment [*molecule\_chembl\_id*, *canonical\_smiles*] (feature selection)
  - h. Compute PaDEL from the selected features (feature engineering)
  - i. Eliminate non-variant features using **VarianceThreshold** method (dimension reduction)



# Feature Engineering and data splitting

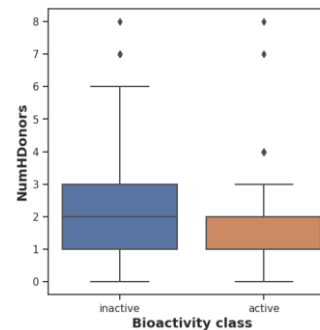
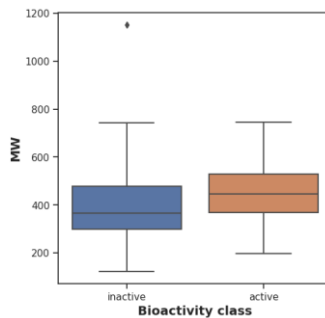
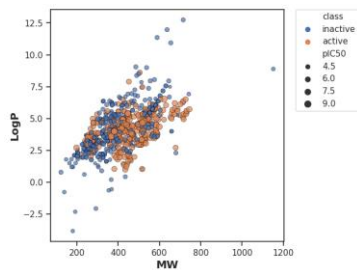
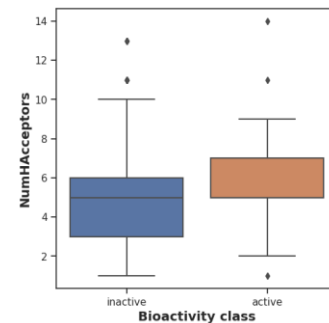
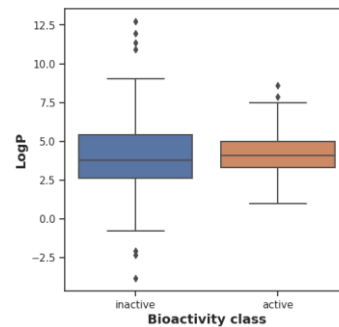
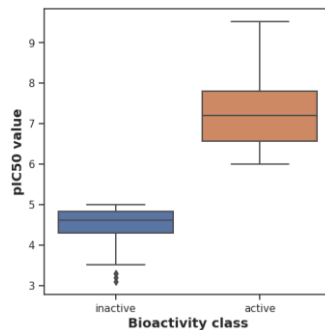
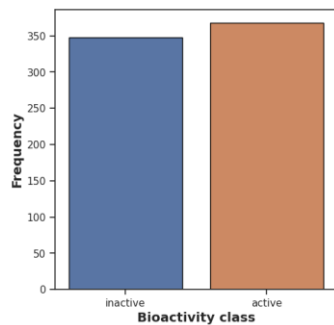
1. Exploratory Data Analysis (EDA)
  - Compute Lipinski descriptor and analyze the dataset in terms of Lipinski criteria
2. Data is split into 67% training 33% testing

	molecule_chembl_id	standard_value	canonical_smiles	class	MW	LogP	NumHDonors	NumHAceptors
0	CHEMBL74874	11000.00	CC12CC[C@@H](C1)C(C)(C)[C@@H]2NS(=O)(=O)c1ccc(...	inactive	327.88	3.83	1.00	2.00
1	CHEMBL75183	10000.00	CC12CC[C@@H](C1)C(C)(C)[C@@H]2NS(=O)(=O)c1ccc(...	inactive	372.33	3.94	1.00	2.00
2	CHEMBL563	305000.00	CC(C(=O)O)c1ccc(-c2ccccc2)c(F)c1	inactive	244.26	3.68	1.00	1.00
3	CHEMBL196279	75000.00	CC(C(=O)O)c1ccc(-c2ccc(Cl)c(Cl)c2)c(F)c1	inactive	313.15	4.99	1.00	1.00
4	CHEMBL195970	77000.00	CC(C(=O)O)c1ccc(-c2cc(Cl)cc(Cl)c2)c(F)c1	inactive	313.15	4.99	1.00	1.00
...	...	...	...	...	...	...	...	...
711	CHEMBL513978	20300.00	CC(C)=CCC/C(C)=C/CC/C(C)=C/Cc1c(O)cc(C)c(C=O)...	inactive	372.51	6.07	3.00	3.00
712	CHEMBL4641877	19900.00	CC(C)=CCC/C(C)=C/CC/C(C)=C/Cc1c(O)cc(O)cc1O	inactive	328.50	6.37	2.00	2.00
713	CHEMBL3609637	31.00	COc1cc(-c2cn(C3CCc4c(F)cccc4N(CC(F)(F)F)C3=O)n...	active	514.48	4.67	0.00	7.00
714	CHEMBL4534005	10.00	COc1cc(-c2cn(C3CCc4cccc4N(CC(F)(F)F)C3=O)nn2)...	active	496.49	4.53	0.00	7.00
715	CHEMBL1091513	0.50	O=S(=O)(NC1CCC(c2cc(F)ccc2F)S(=O)(=O)c2ccc(Cl...	active	517.93	4.67	1.00	4.00

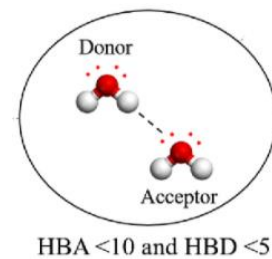
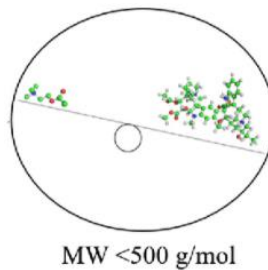
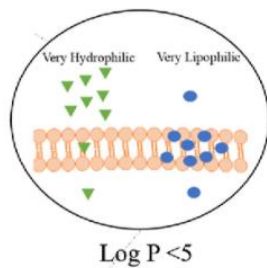
716 rows x 9 columns

# EDA

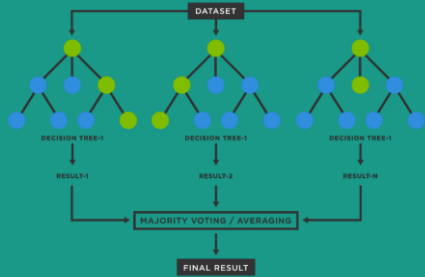
Data Exploration and Visualization using lipinski descriptors attributes



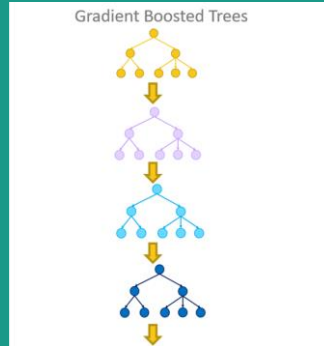
# Lipinski Descriptors



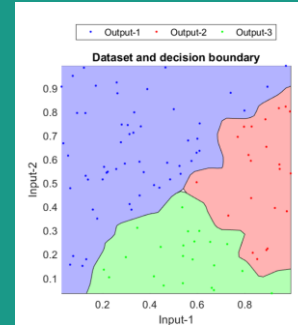
# Data Mining : Regression Problem



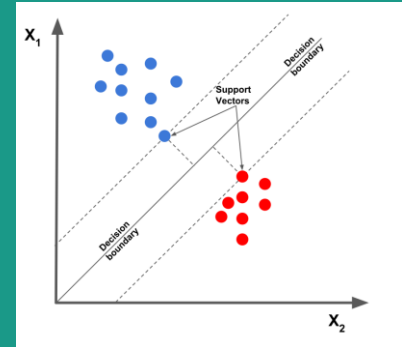
Random Forest



Gradient Boosted  
Regressor



K - Nearest  
Neighbor

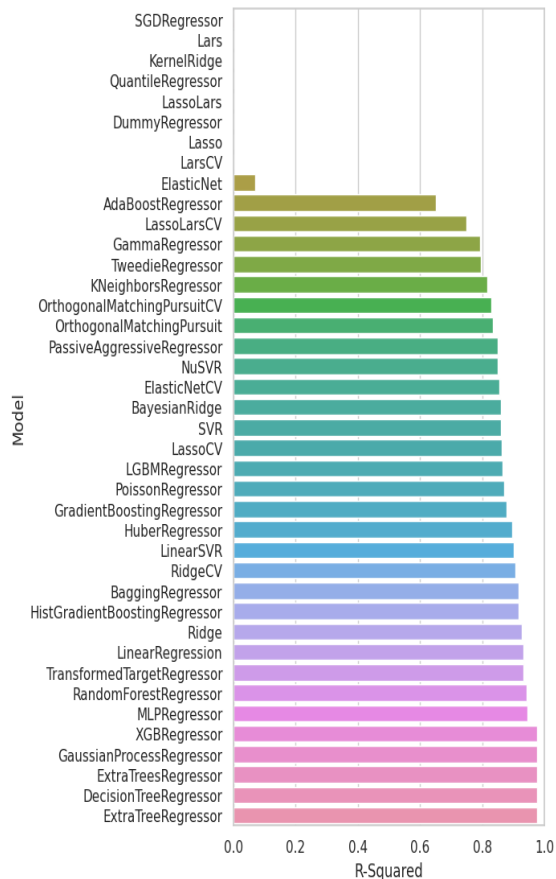


Support Vector  
Machine

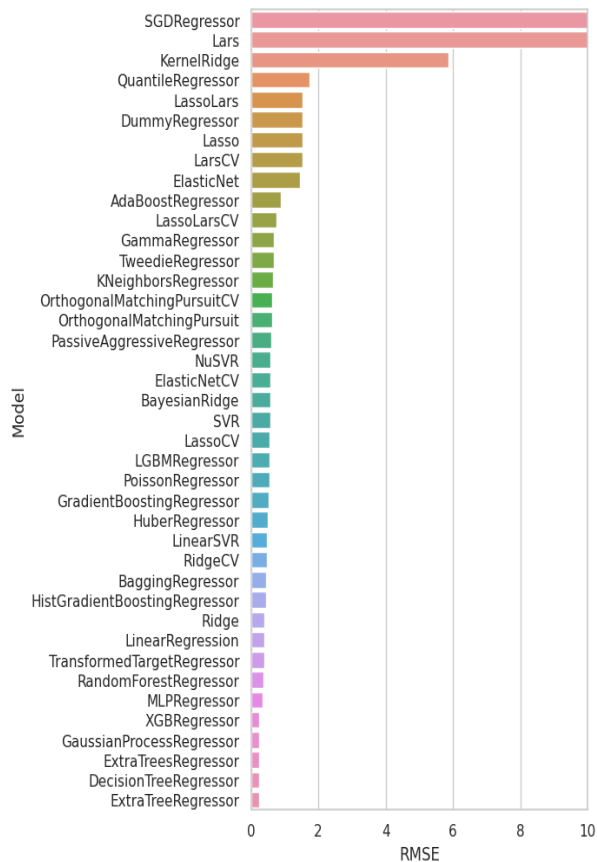
---

# Evaluation of the Lazy Predict Models

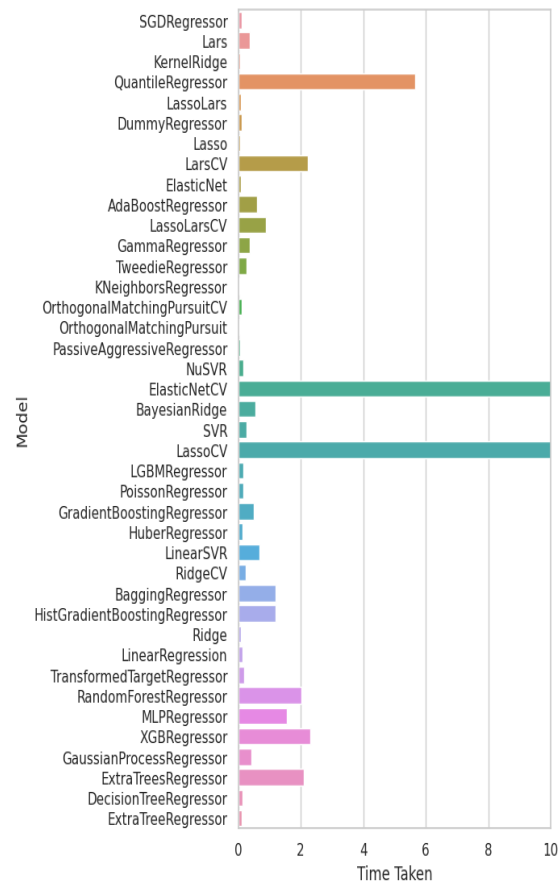
Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
ExtraTreesRegressor	0.96	0.98	0.24	1.08
DecisionTreeRegressor	0.96	0.98	0.24	0.06
ExtraTreeRegressor	0.96	0.98	0.24	0.03
GaussianProcessRegressor	0.96	0.98	0.24	0.13
RandomForestRegressor	0.99	0.93	0.42	0.91
MLPRegressor	0.88	0.92	0.46	1.40
BaggingRegressor	0.83	0.90	0.51	0.12
HistGradientBoostingRegressor	0.81	0.88	0.54	1.66
LGBMRegressor	0.81	0.88	0.54	0.17
XGBRegressor	0.74	0.84	0.64	1.03
GradientBoostingRegressor	0.73	0.84	0.64	0.36
TransformedTargetRegressor	0.66	0.79	0.72	0.03
LinearRegression	0.66	0.79	0.72	0.06
Ridge	0.64	0.78	0.74	0.02
SVR	0.60	0.76	0.78	0.17
KNeighborsRegressor	0.60	0.76	0.78	0.16
NuSVR	0.60	0.76	0.78	0.14
RidgeCV	0.58	0.74	0.80	0.05
HuberRegressor	0.56	0.73	0.83	0.20
SGDRegressor	0.55	0.73	0.83	0.04
LinearSVR	0.52	0.71	0.86	0.25
LassoCV	0.49	0.69	0.89	5.96
BayesianRidge	0.49	0.69	0.89	0.17
ElasticNetCV	0.47	0.68	0.90	7.25
AdaBoostRegressor	0.39	0.63	0.97	0.28
OrthogonalMatchingPursuit	0.30	0.57	1.04	0.03
OrthogonalMatchingPursuitCV	0.30	0.57	1.04	0.06
PassiveAggressiveRegressor	0.09	0.44	1.18	0.03
LassoLarsIC	0.08	0.44	1.19	0.04
LassoLarsCV	0.08	0.44	1.19	0.17



R-squared values



RMSE values



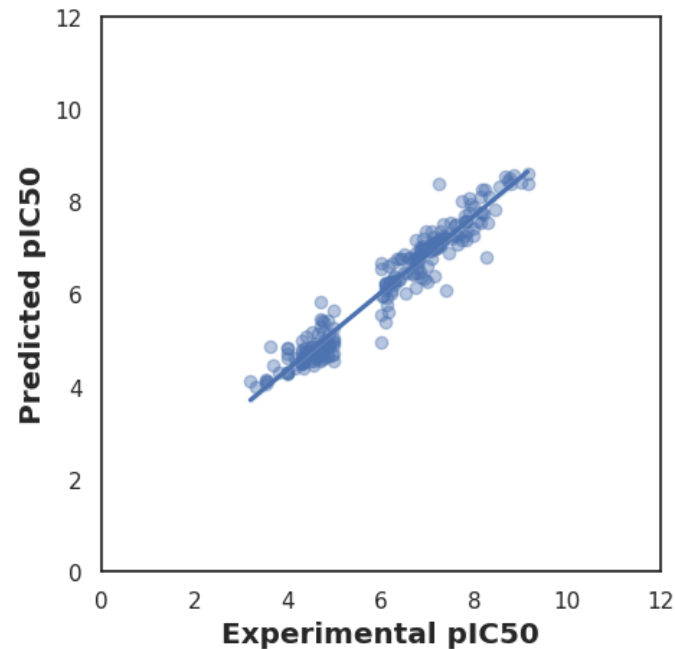
Computation Time



# Hyperparameter Optimization

- Search Grid for the best performing parameters
  - Number of estimators = 800
  - Tree Depth = 8
- Cross Validation , cv = 3

Model	R2 Score	MAE	Execution time (Sec)
Random Forest	0.7045	0.549	0.0091
Gradient Boosted regressor	0.692	0.61	0.0015
K-nearest Neighbor	0.68	0.61	0.0016
Support Vector Machine	0.708	0.61	0.0018
Optimized Random Forest	0.928	0.29	0.0702



---

**Questions & Comments**

**Thank you**